Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 18 (2011) 1–5

Procedia Social and Behavioral Sciences

Kongres Pengajaran dan Pembelajaran UKM, 2010

The Practice of ESL Writing Instructors In Assessing Writing Performance

Wong Fook Fei*, Mohd Sallehhudin Abd Aziz and Thang Siew Ming

Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia

Abstract

The issue of score reliability has always been a contentious one in the testing of language performance because of the subjectivity involved in the assessment process. Assessment of a performance is usually carried out by human raters and studies have proved that there is a lack of consistency and accuracy in such judgments. This leads to a lack of standardization of marks raising concern about fairness to the students taking the course. One way of ensuring reliability is to mandate the use of a language proficiency rating scale. In addition to being a scoring tool, the rating scale also acts as the "de facto construct" (McNamara 1996) and as a term of reference for stakeholders. Despite its importance, its development and use in institutional testing tend to be ad hoc (Fulcher 2008) and hardly ever researched. This paper will report on the preliminary findings of a study that investigates the practices relating to scoring reliability in the assessment of ESL writing. The ultimate aim of the study is to come up with guidelines for improving the reliability of scores awarded for writing assessment.

Selection and/or peer-review under responsibility of Kongres Pengajaran & Pembelajaran UKM, 2010

Keywords: Rating scale; assessing writing; scoring reliability; ESL writing; fairness;

1. Introduction

For a test to be valid, it must test what it purports to test. Also, the rating of the test must be carried out in such a way that the result obtained is reliable. Since the assessment of a direct test of writing is dependent on human raters, who tend to be highly individualistic and idiosyncratic, the scores given for a writing test are often inaccurate, inconsistent and unreliable leading to concerns about fairness to learners.

It is not surprising, therefore, that ensuring the reliability of scores is one of the chief concerns when assessing writing. Research studies carried out mainly before the 1990s, were mainly concerned with coming up with measures to ensure reliability. The measures involved improvement and refinement in rating scale design and use, and standardization procedures such as rater training and moderation sessions (for example Ruth and Murphy: 1988, Weigle: 2002 etc). These studies have shown that reliability of scoring writing performance can be attained through the use of rating scale which specifies the requirements of a writing task in the forms of descriptors (Alderson 1991, Hamp-Lyons 1991) and training session where raters are taught to use the rating scale by looking at benchmark

* Corresponding author. Tel.: +6-012-662-6739; fax: +603-8925-4577

E-mail address: wff@ukm.my.

1877–0428 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

doi:10.1016/j.sbspro.2011.05.001

samples. Such reliability methods have been adopted in most rating procedures involving direct assessment of writing and have been included in language testing references written for practitioners such as Hughes (2003), Weigle (2002) and Weir (2005).

1.1. Ensuring Scoring Reliability

Most textbooks on language testing will provide some guidelines on how to ensure scoring reliability. Hughes (2003:94-107), for example, suggests the following measures in order to ensure the reliability of scores awarded for writing performances:

- Create appropriate scale for scoring
- Calibrate the scale to be used
- Select and train scorers
- Follow acceptable scoring procedures

Hughes stresses on the design and calibration of the rating scale and the training and moderation of the raters. In Weir (2005)'s socio-cognitive framework for validating skills in language tests, reliability (termed scoring validity) is considered a part of validity and not as a separate entity. The four main components that are involved in the rating process are criteria or rating scale, rating procedure, raters and grading and awarding. Under the rating procedure, Weir lists variables like rater training, standardization, rating conditions, rating, moderation and statistical analysis of the results. Weir argues that the more evidence that is collected the stronger the validity of a test. The higher the stakes of the test, the more stringent would be the demand for evidence of each of these components.

This study is framed using Weigle's (2002) approach to scoring reliability and seeks to investigate the practices of scoring ESL writing by examining the rating scales used and the rating procedures of undergraduate English proficiency courses that develop writing skill. For this paper, the practice of one course is discussed.

2. The Study

This study is conceived because of a perceived lack in the one aspect of the testing procedure in UKM. Quality assurance in testing at UKM is concerned mainly with the guidelines for the conduct of examination, and with the form and content of the examination paper, which is moderated by a vetting panel formed by the school. The scoring procedure, on the other hand, does not receive as much attention. Here, what is mandated is the existence of a marking scheme and a score sheet for the entry of marks. For a writing test, this would mean that there should be an assessment scale which will guide the raters in the evaluation process. However, there is no prescribed procedure for ensuring scoring reliability of students' written scripts. Such guidelines would ensure that inaccuracy does not arise from intra and inter-rater unreliabilities, and that a score given to a student accurately reflects his achievement and ability as displayed in the course. In order to come up with such guideline, it is necessary to examine the current scoring practice.

The details about the study are briefly described below.

2.1. The Purpose

The study examines the scoring practices involving ESL writing assessment as carried out in an English proficiency course. It seeks to find answers to the following questions:

- What rating scales are used for assessing and how were the rating scales developed?
- What procedures have been put in place to ensure that the rating scales are adhered to when rating?

2.2. The course

The course that is the focus of this paper is an advanced ESP (English for Specific Purposes) course that equipped students for demands of English in the workplace. The course booklet describes the aim of the course as

"to equip students with both local and written communication skills which prepare them to perform well in various workplace situations". The course basically focuses on the two performative skills of writing and speaking. The course is taken mainly by Engineering students, and students who have completed the prescribed EAP (English for Academic Purposes) courses.

In terms of writing, the two learning outcomes as listed in the course booklet are:

- Ability to write effective resume and cover letter
- Ability to write long analytical reports

The two writing products that are assessed arising from the two learning outcomes are – 1) resume and cover letter, and 2) report. There is no summative assessment and the two writing products that are assessed constitute the writing component of the wholly formative assessment components of the course.

2.3. The Method

Data presented here are obtained from an interview with the course coordinator and from an examination of the course booklet and rating scales used. The interview questions were focused on the 2 scoring variables: the design and use of the rating scales and the rating procedures that ensure inter rater reliability.

3. Discussion Of Findings

This preliminary report focuses only on the analysis of the data from two sources, the interview with the coordinator and examination of the course booklet which contains all relevant information concerning the conduct of the course. The later development of the research would involve interviews with other course coordinators and examination of more rating scales and rating procedures.

The following is a discussion of the main findings.

3.1. Rating Scale Design

An analytic rating scale is used for assessing the written products of each of the two tasks. The scales are adopted and adapted from rating scales previously used for similar courses. The scales have been in use for a quite a while and have evolved from previous courses. The instructors are thus familiar with the scales and their ease of use has been established. The selection of rating scales appears to be dependent on the experience of the course committee and the kind of rating scales that they have been exposed to, and have experience working with.

Analytic scoring is more suitable for the course as the tasks are formative in nature. Also, diagnostic details about areas of weaknesses could be provided to the students so that they could do more focused remedial exercises. This is in line with the learning focus of the assessment procedures of the course. In addition, analytic scoring is considered more appropriate for ESL writing which has been found to be uneven across the different sub-components of writing (refer to Hamp-Lyons, 1991)

There are four rating criteria: format, language, content and overall impression. The descriptors for format are specific to the task while the other criteria are more broadly defined and thus similar for both writing tasks. The three generic criteria are the same as the ones used in MUET (Malaysia University English Test), a mandatory English test for entrance to local institutions of higher learning. The mirroring of criteria from MUET can be construed as a positive one as there is a need for a common framework of English language Proficiency so that standards can be normalized for all English courses offered at institutions of higher learning regardless of courses and institutions offering them.

The ESL construct can be derived from the specifics of the rating scales. The descriptions of the criteria and the weighting given informed that appropriateness and relevance of content and language are of particular importance while format to define the genre is also emphasized. The workplace proficiency appears to be demonstrated by the format and is considered in terms of relevance and appropriateness.

3.2. Rating Scale Use

While there are rating scales to guide instructors when awarding marks, the questions that should be asked are 1) whether there is a common interpretation of the scales among the instructors and 2) whether the rating scales are actually referred to and when scoring.

No rater training was conducted to help instructors interpret the rating scales and to standardize the ratings. Instead, a briefing was conducted before the commencement of the course where general information about the course was given. Here, instructors were briefed about the evaluation procedure of the course. To help instructors come to grip with the requirements of the tasks and the standard expected, samples of previous students' cover letters and resumes, and reports that had been benchmarked were made available for reference. The onus was on the instructors to look through these samples and to conform to the standards that had been predetermined.

The general feeling was that training was not necessary as the instructors were already familiar with the rating scales as they had been in use for a while and the instructors are all experienced. Also, it was felt that such training session would be regarded as supercilious by the instructors who would rather go about their daily activities of teaching with as little interference as possible. The feeling was that the instructors should be trusted to go about their duties.

Lastly, there was also no moderation of the scores given by the various instructors to ensure that judgment of quality is consistent across the different classes. The belief is that the existence of a rating scale which specified how marks are to be awarded is sufficient, and that experienced instructors should be able to judge the writing quality of their students and that everyone's notion of quality as described would be the same. While there were concerns about lenient instructors who gave very high marks to his/her class, or who award marks for effort rather than the quality of the products, these have largely been brushed aside. There is a general reluctance to question the professionalism and integrity of a follow colleague so scores have always been accepted as presented with no counter-checking. The tacit agreement is that the class teacher is in a best position to judge the performance of his/her class.

Obviously, such practices compromised the scoring validity of the course. In such a scenario, the score that a student receives is not comparable between the different classes taught by different instructors. This can be seen as trade off as the assessment is formative and not summative and the focus is learning, albeit at the expense of scoring reliability.

To summarize, scoring procedures are largely ignored as teacher classroom autonomy is deemed more important.

4. Conclusion

The preliminary findings indicate that procedures to ensure inter-rater reliability are lacking leading to problems with comparability of scores across the different classes. Except for the existence of rating scales, there is hardly any scoring procedure to ensure consistency in awarding marks. Granted that the course is formative and focused on learning but we also need to be fair to students who want to receive a fair grade for their effort.

In view of the findings, it is suggested that two procedures - premarking session and submission of marked scripts - should be implemented for writing courses taught by a team of instructors.

- Pre-marking session should be held

Attendance for all instructors should be compulsory for both experienced and novice raters. Benchmarked scripts from the current cohort of students should be used for trial marking. Adjustment could be made to the descriptors in the scale if necessary so that it better reflects actual performance. The scores given to these scripts should be agreed to by all. This is necessary to ensure that everyone has the same interpretation of the scale and knows what features to look out for when assessing. It is also important for standardization purposes.

- Submission of sample marked scripts for each grade level

Instructors should be asked to submit an excellent, a good, average and poor script from his/her class so that the coordinator can double-check that the rater has kept to the standard set. If the rater has been lenient or too strict with his/her marking, then this has to be pointed out and adjustments made to his/her class marks.

With the imposition of these two measures, there would be check and balance between marks given by the various instructors in the course.

5. Acknowledgements

This research is funded by a university research grant, Geran Penyelidikan Tindakan (Code UKM-PTS-061-2010) provided by Universiti Kebangsaan Malaysia.

References

Alderson, J.C. (1991). Bands and scores In J.C. Alderson, and B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp.71 86). London: Modern English Publications/British Council. Macmillan.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment an advanced resource book*. Routledge: London.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.). *Assessing second language writing in academic contexts* (pp.241-276). Norwood, N.J.: Ablex Pub. Corp.

Hughes, A. (2003). *Testing for language teachers*. (2nd ed.). Cambridge: Cambridge University Press.

McNamara, T.F. (1996). *Measuring second language performance*. London: Longman.

Ruth, L. and Murphy, S. 1988. *Designing writing tasks for the assessment of writing*. Norwood, N.J: Ablex.

Weir, C.J. (2005). *Language Testing and Validation An evidence-based approach*. London: Palgrave Macmillan

Weigle, S.C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.